CareerFoundry Immersion Task 6.1

Andre Nedvidek

Data Source

I have selected a [USA Real Estate Dataset](https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset), sourced from a reputable real estate company called ‘Realtor’. This company is well-known, ensuring the reliability of the data.

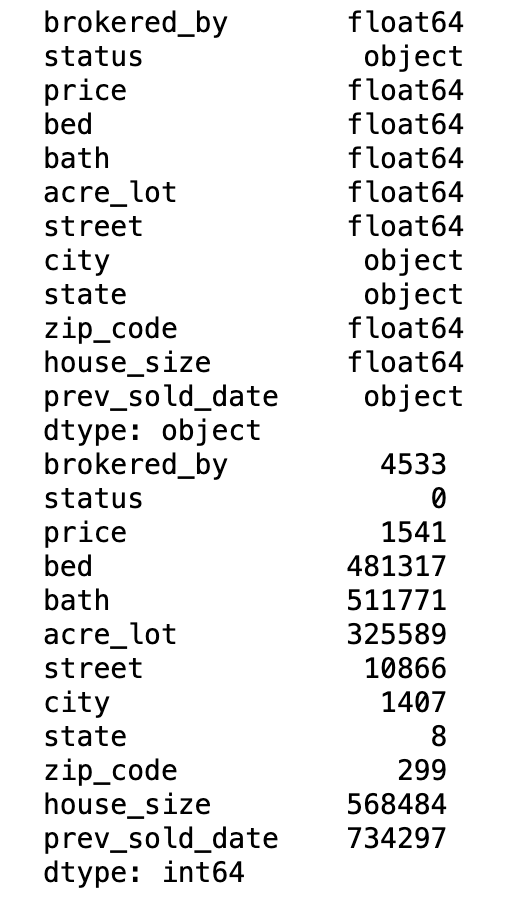
The dataset meets all the criteria specified in the Project Brief. It contains over 1500 rows of continuous and categorical data and is well-suited to fulfill the project's analytical requirements.

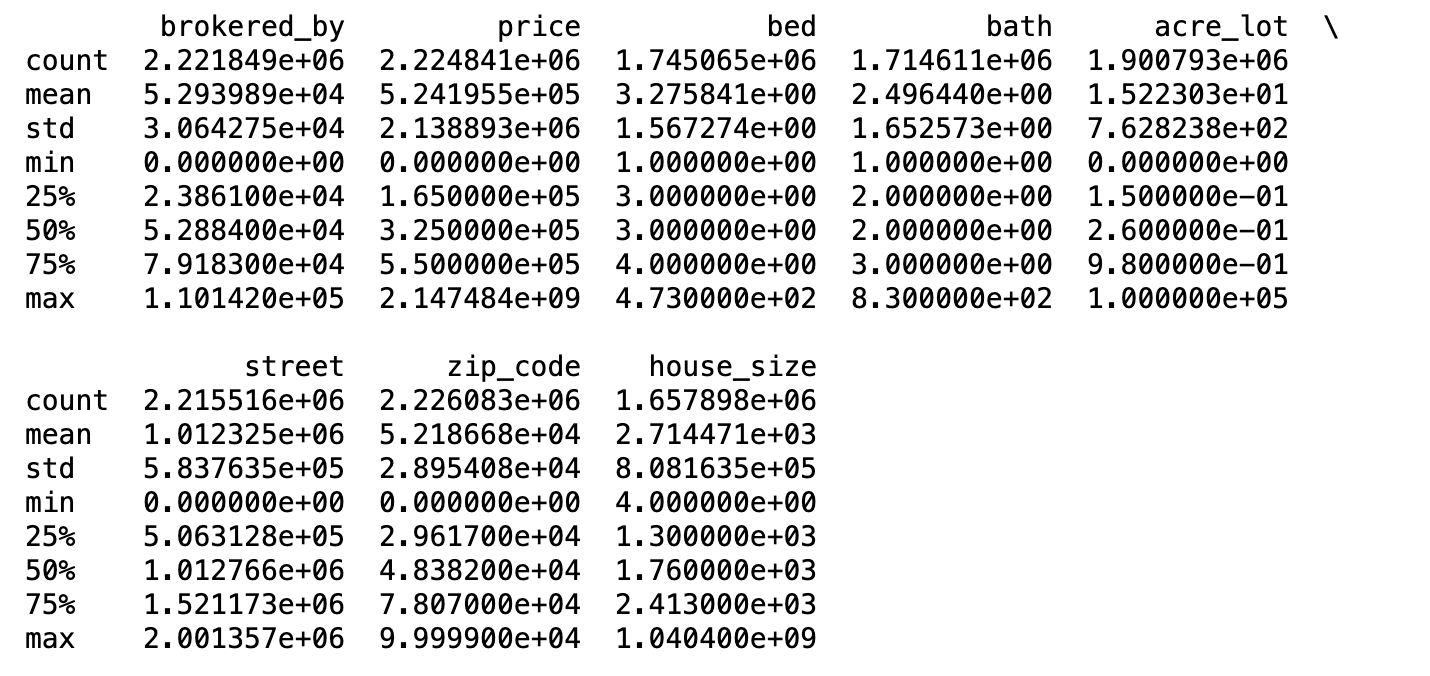
The dataset includes historical housing price data across the USA, along with details such as square footage, number of bathrooms, and more.

Data Choice

After an extensive search, I selected this dataset as a backup option, despite it not initially being my preferred topic. Upon closer examination, I've discovered its untapped potential and now feel confident in crafting a compelling narrative from the data

A screenshot of a graph

Description automatically generatedData Profile



Original Size:

12 Columns, 2226382 rows

Data Cleaning:

1. Dropped Column ‘Brokered-By ’
2. No Duplicates were found
3. Deleted Rows with Missing Values for ‘Price’, ‘city’ and ‘state’

New Data Set Size:

11 Columns, 2223429 rows

**Consider limitations and ethics.**

**The dataset presents a few limitations and ethical considerations. The unclear time range and potential geographical bias may affect the data's relevance and representativeness. These factors should be clearly acknowledged in any analysis to ensure accurate interpretation of results.**

**Ethically, privacy remains a concern despite encoding efforts, as combining location data with specific house attributes could risk re-identification. The dataset may reflect and potentially perpetuate societal biases present in the housing market. Using this data for predictive modeling raises fairness concerns, especially given historical inequities in housing. Additionally, the data's restricted educational use, lack of transparency in processing, and unclear consent from data subjects present further ethical challenges.**

**To address these issues, researchers should clearly state the dataset's limitations, avoid broad generalizations, and consider the real-world implications of their analyses.**

**Questions to explore**

Geographical Price Variations:

How do median housing prices vary across different states and major metropolitan areas?

Property Size Distribution:

Analyze the distribution of property sizes across different regions.

Categorical Predictors of Housing Prices:

Which categorical variables (e.g., property type, number of bedrooms, neighborhood) are the strongest predictors of housing prices?

Property Clustering by Amenities:

Perform a k-means clustering analysis on properties based on their amenities (bedrooms, bathrooms, square footage).